Andrew Nickeson and Leah Sohn

Lab Journal Questions 1

1. What does it mean to “Tidy” your data?

Data does not come prepackaged to our own purposes. Typically, the data needs to be stored in a consistent way which allows you to work with it properly. Every column is a variable, and every row is an observation.

1. What does it mean to “Transform” your data?

Transforming data primarily involves filtering the data down to the variables which concern the visualization you are pursuing. This sometimes involves creating new variables which are themselves functions of previously existing rows and columns in the data.

1. How important is the “data visualization” skill for your career?

Data visualization usefulness is entirely situational. There are situations in most careers where it is certainly very helpful. I currently don’t have a career nailed down, but as a student I find that tracking my workload is extremely helpful. I haven’t visualized the data, but I do try to keep track of previous trends in workload (tracked in hours) and try to make predictions on future workloads based off of this data. These predictions would be made easier with data visualization. I also run a lot for my sport, which sometimes can feel like a job. I have very often made data visualizations of my mileage to see where I’ve been and pick out some patterns that may lead to problems.



Map

Description automatically generated<https://shiny.rstudio.com/gallery/india-blood-banks.html>

This data visualization uses a data set which stores blood bank locations and their affiliation (charity, private, government). It then places these points on a map and color codes them based on affiliation (green for charity, red for government, purple/blue ish for private). As stated in the read me file, the data is not 100% accurate and there are “close to 82 NA values for latitude and longitude” meaning that those banks couldn’t be placed on the map. Affiliation data is purely categorical and is used only for the aesthetic marking. Otherwise we are dealing with quantitative data in the latitude and longitude of the various blood banks which have been mapped.

This mapping didn’t use any real aesthetic markers on the axes, but it did on the points in order to color the points. It would have also had to have used coord\_quickmap() in order to set the aspect ration correctly to have the points placed correctly on the map.

1. What is the difference between the following command lines:

git config user.email “[email@example.com](mailto:email@example.com)”

or

git config --global user.email "email@example.com"

The difference is the --global flag which allows us to write to our global git configuration. Without the --global flag we only change the configuration of the repository. In the above case we are adding our account email to our machine overall with the --global flag. However, if we wanted to use a different git account email for a different repository (like say if someone else were to use your machine), then we would drop the flag.

1. What are the five key dplyr functions described in the chapter 5?

The five key dply functions are filter(), arrange(), select(), mutate(), and summarise().

1. What are cumulative and rolling aggregates?

A cumulative aggregate would use all of the data provided, so if you gave a list of numbers and got the cumulative sum, you would just have the sum of all the numbers. The rolling aggregate is going to take into account only the values listed previously as it progressively rolls through the data. So it will sum all the numbers prior to where its at, and only at the end of the sequence will it sum all of the numbers.

1. What are ranking functions?

Ranking functions will return the position of the number in the sorted vector, but it will leave the ordering of the actual numbers alone.

1. Imaging that you are already assigned to work with a team, what does that team need to do to make that project a success? Provide concrete examples.

The team needs to communicate effectively with each other and the client in order to create a plan which everyone understands and agrees upon. This ensures that everybody is starting on the same page. It is likely that the plan will change, so effective communication throughout is important. Sometimes priorities of the project change and sometimes the team runs into an unforeseen issue. The plan will need to change.

1. What is the most important difference between generic software product development and custom software development? What might this mean in practice for users of generic software products?

Generic software development entails that the organization developing the software gets to control where the software goes when it runs into problems. They get to set the specifications. Whereas for custom development the developer works solely to meet the requirements of the customer. If the project runs into a problem, they must consult the customer and the customer decides which direction to go. For generis software product users this means a less specialized and more general experiences. Not all their most important needs are going to be met most of the time. The custom development entails a far more specific and tailored experience.